Data Statistics Theory

Kristhian Santiago Palomino Fajardo

May 2024

1 Introduction

This document provides an overview of the statistical measures used for data analysis, including mean, median, trimmed mean, mean absolute deviation, standard deviation, and quantiles. These measures help in understanding the central tendency, dispersion, and distribution of the data.

2 Mean

The mean (or arithmetic mean) is the sum of all values divided by the number of values. It is a measure of central tendency.

$$Mean = \frac{\sum x_i}{n} \tag{1}$$

where x_i are the data values and n is the number of data points.

3 Median

The median is the middle value when the data points are arranged in ascending order. If the number of data points is even, the median is the average of the two middle values. It is less affected by outliers and skewed data.

4 Trimmed Mean

The trimmed mean is a measure of central tendency that removes a specified percentage of the smallest and largest values before calculating the mean. This reduces the effect of outliers.

$$TrimmedMean = \frac{\sum_{i=k+1}^{n-k} x_i}{n-2k} \tag{2}$$

where k is the number of values removed from each end of the data set.

5 Mean Absolute Deviation

The mean absolute deviation (MAD) measures the average distance between each data point and the mean. It provides an idea of the variability in the data.

$$MAD = \frac{\sum |x_i - Mean|}{n} \tag{3}$$

6 Standard Deviation

The standard deviation is a measure of the dispersion or spread of data points around the mean. It is the square root of the variance.

$$StandardDeviation = \sqrt{\frac{\sum (x_i - Mean)^2}{n-1}} \tag{4}$$

7 Quantiles

Quantiles are points in the data that divide the data into equal-sized intervals. The most commonly used quantiles are the quartiles, which divide the data into four parts:

- **25% Quantile (First Quartile, Q1)**: The value below which 25% of the data fall.
- **50% Quantile (Median)**: The value below which 50% of the data fall.
- **75% Quantile (Third Quartile, Q3)**: The value below which 75% of the data fall.

8 Boxplot

A boxplot is a graphical representation of the distribution of data based on a five-number summary: minimum, first quartile (Q1), median, third quartile (Q3), and maximum. It helps in identifying outliers and the spread of the data.

9 Histogram

A histogram is a graphical representation of the distribution of numerical data. It shows the frequency of data points within specified ranges (bins).